Crowdsourcing Linked Data on listening experiences through reuse and enhancement of library data

International Journal on Digital Libraries
https://doi.org/10.1007/s00799-018-0235-0

Crowdsourcing Linked Data on listening experiences through reuse and enhancement of library data

Alessandro Adamou[1] · **Simon Brown**[3] · **Helen Barlow**[2] · **Carlo Allocca**[4] · **Mathieu d'Aquin**[1]

Received: 18 August 2016 / Revised: 19 December 2017 / Accepted: 18 January 2018

Abstract
Research has approached the practice of musical reception in a multitude of ways, such as the analysis of professional critique, sales figures and psychological processes activated by the act of listening. Studies in the Humanities, on the other hand, have been hindered by the lack of structured evidence of actual experiences of listening as reported by the listeners themselves, a concern that was voiced since the early Web era. It was however assumed that such evidence existed, albeit in pure textual form, but could not be leveraged until it was digitised and aggregated. The Listening Experience Database (LED) responds to this research need by providing a centralised hub for evidence of listening in the literature. Not only does LED support search and reuse across nearly 10,000 records, but it also provides machine-readable structured data of the knowledge around the contexts of listening. To take advantage of the mass of formal knowledge that already exists on the Web concerning these contexts, the entire framework adopts Linked Data principles and technologies. This also allows LED to directly reuse open data from the British Library for the source documentation that is already published. Reused data are re-published as open data with enhancements obtained by expanding over the model of the original data, such as the partitioning of published books and collections into individual stand-alone documents. The database was populated through crowdsourcing and seamlessly incorporates data reuse from the very early data entry phases. As the sources of the evidence often contain vague, fragmentary of uncertain information, facilities were put in place to generate structured data out of such fuzziness. Alongside elaborating on these functionalities, this article provides insights into the most recent features of the latest instalment of the dataset and portal, such as the interlinking with the MusicBrainz database, the relaxation of geographical input constraints through text mining, and the plotting of key locations in an interactive geographical browser.

Keywords Linked Data · Musical reception · Crowdsourcing · Music history · Text mining · Bibliography

1 Introduction

Listening to music is by its own right a notion that can be, and is indeed being, approached by the research community in a multitude of ways. Several research strands targeting the listening experience have been detected in social, cognitive and economic sciences: the reception of music is analysed from the viewpoints of either elite audiences, as in evidence of music criticism, or the masses, if almost exclusively through commercial indicators of contemporary popular music. The psychological dynamics of the listening activity per se [21,31,33], and its effect on human behaviour in everyday contexts [22], are also being extensively investigated.

The extent of these research efforts is necessarily tied to the nature and availability of evidence, as well as the representational capabilities of the devices through which the

✉ Alessandro Adamou
alessandro.adamou@nuigalway.ie

Simon Brown
simon.brown@rcm.ac.uk

Helen Barlow
helen.barlow@open.ac.uk

Carlo Allocca
c.allocca@samsung.com

Mathieu d'Aquin
mathieu.daquin@nuigalway.ie

1 National University of Ireland Galway, Galway, Ireland

2 The Open University, Milton Keynes, UK

3 Royal College of Music, London, UK

4 Samsung Inc., London, UK

Published online: 06 February 2018

evidence is presented. Whilst the existing investigation on music reception can count on sales figures, ratings, published professional reviews and some user studies, these are only part of the data that could drive a comprehensive study. As early as the 1990s, special editions of top journals in musicology had drawn attention to the difficulties of identifying primary source material about experiences reported by actual listeners, and of collating them in quantity [11,27,38], in order to provide a robust evidential basis which would underpin such a field of study. One reason behind this gap may be the combination of a representational problem (i.e. obtaining the evidence as data that can be machine-read aggregated and analysed), practical or legal difficulties in obtaining access to the sources, and lack of resources (hardware, trained manpower, etc.) for implementing the required transformations. Issues with the accuracy and completeness of the evidence may also get in the way of giving a structured representation to this information as it becomes machine-readable data.

One way to look at how listening to music has been experienced by individuals over the course of history is through an approach that is equally music-centric and bibliographical: there is indeed a strong bibliographical element in representing evidence from personal correspondence, chronicles, recordings from third parties and other sources. If on the one hand the representation has to allow for vague, uncertain or ambiguous evidence, on the other hand the sources of evidence should be formally represented so as to highlight their relations to other sources and elements of a musical experience. Indeed, not every online digital library and bibliographical dataset goes the extra mile of providing linkage between a letter or diary entry and the collections where it appears, or an excerpt within it.

The **Listening Experience Database (LED)**[1] is a project conceived to respond to this challenge through the application of crowdsourcing methods and data technologies to enable the creation, representation and publication of structured data about listening experiences. Indeed, though the required evidence may only sporadically be present on the Web, a great deal of information that revolves around it, concerning for instance writers, musicians, published books, places and musical works, already exists online, freely and in structured form.

To bootstrap the database with this knowledge already embedded, the principles and technologies known as *Linked Data* (LD) [7] were adopted in LED. With this method (whose staples include reusing URIs to name all things, storing and publishing in the RDF format and querying in the SPARQL language), it is possible for LED to use external open data to enrich the human-readable version of its content, through its Web portal, and also to publish enhancements of the original data. LED uses Linked Data natively, in that its data are internally stored in RDF using the same model as the one used for publishing them, thus eliminating the need for transformations across traditional database systems. In addition, the input model adopted by LED is that of *crowdsourcing*, where the systematic accumulation of primary data takes place at the hands of a community of scholars and enthusiasts, who received minimum guidance as to the data entry process itself. Linked Data is an "open-world" paradigm—so not all the data about something need to be entered together in one place at one time—but it is also rigorous with respect to the identity of things: every referenceable thing, such as a person, book, musical composition, performance, or text excerpt, should be uniquely identifiable and described. Dealing seamlessly with the identities of things and hiding this complexity from the crowdsourcing community, allowing them as much liberty as possible to enter even vague, under-represented and uncertain information, was a major challenge in the project. Another challenge that is purely technological concerns the capability of Linked Data technologies to manage the data-intensive traffic of such a database natively and in real time. The representational choices and the shape of the data being stored and managed had significant repercussions on the effectiveness and efficiency of user interaction, which prompted trade-offs to be sought, which in turn had an impact on the stored representation.

Whereas the central part of this paper will describe the data platform and its underlying model, the later part will focus on the choices that were made to tackle the aforementioned data management issues.

2 Related work

Key to any study on the contexts of music listeners is the extensive and articulated representation of what is being listened to, as well as of who is listening and where/how, each to various degrees and facets. Whilst to the best of our knowledge no previous effort tried to establish a research discipline that equally encompasses all these aspects, several recent lines of research were found to concentrate on pairs of them. *Gracenote* recently conducted a study on identifying relationships between music tracks and listening contexts associated with public holidays [35], while Cunningham et al. relate the social and auditory aspects of listening during road trips [14]. Thanks to the abundance of models and data for multimedia [6], other studies were able to concentrate on formally modelling musical mood and its relationship with specific audio features [37]. On relating the "who"s and "what"s of the listening experience, Schedl et al. [32] characterise listeners by background and personality traits and measure their agreement on the subjective perception of

[1] Online at http://listeningexperience.org.

classical music being played, such as the perceived instruments and tempo.

With the recent years showing an empowerment of the everyday user as an influencer over the global sentiment on music alongside the professional critic, later studies have begun to acknowledge the notion of *amateur critique* through activity on `genius.com` [15] and other sound annotation platforms.

Crowdsourcing accounts of personal experiences in the literature and throughout history is not an entirely new practice in research in the Humanities. As a matter of fact, an earlier sister project to LED called the **Reading Experience Database** (RED) [8] aimed at capturing evidence of reading rather than listening, thus resulting in about 34,000 records, of which about one-fifth was contributed by the community. This project corroborated the hypothesis behind LED concerning the availability of unassembled evidence about personal experiences in the literature. As RED was built on top of a traditional database system that did not contemplate external data reuse until a final step of translation to Linked Data, LED set out to include this aspect in each phase of its data management workflow, thereby opting for the development of a native Linked Data framework.

Owing to the popularity of the domain among people and online communities, the practice of crowdsourcing musical data through Wikis and other platforms is widespread throughout the Web, as demonstrated by portals such as *Discogs*[2] for discographies and *Setlist.fm*[3] for concert setlists. Many of these databases thrive on the large availability of information among the masses, drawing from the very own experiences of users—their record collections and concerts attended—and therefore mostly restricting to the phonographic era. Therefore, the only possible overlap with LED is limited to documented accounts of contemporary history. Additionally, these databases tend to be generated in isolation from others, with little to no interoperability with other sources at data creation time.

Very recent research endeavours have postulated the need for a *crowdsourcing Linked Data* discipline by its own right which, in its present state, can still be considered in its infancy. In fact, most of the relevant applications that can be cited belong in the initial running period of LED. Particularly, some work has explored the lowest layers of crowdsourcing Linked Data, namely in the definition of tasks and the extent to which they affect the life cycle of data [4]. This contribution stems from some foundational work by Simperl et al. [34], which addressed the definition of a conceptual architecture for crowdsourcing LD and the relationships between task definitions and SPARQL queries.

Further of interest to LED is a most recent application of crowdsourcing in order to spot and correct factual errors in the data management by a system [30]. However, it has not been reported whether any form of data multi-tenancy is in place, which would allow moderators to detect whether a suggested correction is supported by the community; a feature that was felt as required for the LED data workflow. We also record attempts at tackling crowdsourced linked data in the mobile domain, such as in the mobile framework **CLODA** [23]; however, these systems rely on the (mostly non-conscious) aggregation of input from the devices of the community, which raised concerns as to what data curation and validation processes are employed.

Since LED, other significant research efforts have been striving to consolidate the formal structuring of the listening practice. **DOREMUS**[4] is a project targeting models and controlled vocabularies for publishing collections of musical metadata [1]. Whereas the focus of LED is on off-the-shelf reuse of data, thereby importing, where possible, the most common ontologies and vocabularies they are represented with, DOREMUS aims at providing guidelines for generating reusable data in the first place, top-down and through adaptive vocabularies. One of their early results is the extension of the *FRBRoo* model [5] for the representation of both musical works and musical events. Possible synergies with the DOREMUS data model and avenues for backporting them to LED are currently under investigation.

3 LED overview

The Listening Experience Database is a twofold contribution in the forms of (1) a semantically enriched, machine-readable dataset about musical events, records of experiences of listeners, their source documents and all the actors involved, and (2) an online data management platform that adds linked data support to a content management system (CMS) and makes this dataset directly available for faceted searching and browsing—also geographical. It has received submissions for nearly 10,000 recorded accounts of listening to music throughout history; of these, over 90% have so far undergone validation and are publicly available, with the remainder pending review by moderators in the project team. Submissions were contributed by a team of 50 users comprising project members, scholars and enthusiasts in music and musicology. The source documentation of the evidence in text for these records consists of about 300 primary sources between published books and collections of as-yet unpublished manuscripts.

[2] Discogs, http://discogs.com.

[3] Setlist.fm, http://setlist.fm.

[4] DOnnées en REutilisation pour la Musique en fonction des USages, http://www.doremus.org.

3.1 Inclusion protocols

In compliance with the protocols established by the project for the data to be captured,[5] the dataset has a wide span with regard to most dimensions of the listening experience context that were considered for formal representation (cf. Sect. 4 for details on the semantics of LED data). Particularly, listening experiences may have been recorded in any historical period, anywhere and for music of any genre denomination; in fact, "music" is here intended as including any form of organised sounds, written or unwritten. The listeners whose experience is recorded may coincide with or differ from the authors of the evidence, and the experiences may be deferred in time and space with respect to the musical performances. Also, none of this information is required to be known at input time.

Whereas the factual elements of the listening experience context are intentionally lax, the inclusion protocols place constraints on the allowed source documentation, in order to guarantee the respect of basic scholarly conventions. The categories of allowed source material include published books, letters, diary entries, news, scientific or state papers and even social network entries, and are subject to the following constraints:

- The source material need not be officially published at data entry time: manuscripts are allowed.
- It must not be original, i.e. it must exist in writing prior to being entered into LED.
- It must exist in English language, whether in its native form or officially translated.
- It must not be a work of fiction or solicited criticism, such as an album review or live report on a professional magazine, as these are outside the project's remit.

These protocols also provide the specification for minor constraints, such as the syntax for labelling sources that do not have an official title, like individual letters and diary entries.

3.1.1 Governance policies

The governance model of content creation in LED is one of *supervised crowdsourcing*. Any user may request a role as a contributor in LED and has no limit to the amount of content and details they can submit to the system, within the aforementioned inclusion protocols and scholarly conventions. However, in order for submitted contributions to be included in the public dataset, they require approval by a team of moderators comprised of or designated by the project team.

Crowdsourcing in LED was modelled around *macrotasks* of controlled complexity [12]. The basic macrotask is to input the description of a listening experience in its entirety, including the data that describe the source documents, people, musical works and performances thereof. If any of these peripheral entities already exist in the dataset, they can be reused, in which case their descriptions need not be re-entered; however, users may edit them, which is equivalent to proposing amendments that also need to be validated before they are propagated to the public dataset. Over the course of the crowdsourcing phase, the granularity of these macrotasks has proven to be manageable enough for mildly experienced users to be able to complete one in as little as five minutes without negatively affecting the quality of their data.

Support for *microtasks* such as providing or editing data on another individual than a listening experience (e.g. proposing a change to the date of birth of a musician or the URL of a source) was ruled out at protocol definition time in order to prioritise the entry of unique content; proposals of such amendments have only been possible through the entry of a new listening experience record that reuses these entities. However, microtasks are being considered for introduction, having the amount of listening experiences now reached the required critical mass.

3.2 Dataset

The machine-readable LED dataset is one of the RDF graphs published on the linked data platform of the Open University;[6] its graph is named http://data.open.ac.uk/context/led. At the time of writing, the graph contains 416,818 RDF triples, describing 9019 approved records of listening experiences out of approximately ten thousand total submissions. It is also the fourth largest RDF dataset out of 36 published on that platform,[7] and the largest to be the output of a research project. Once the remaining submitted records pass the approval phase, it is estimated that the graph size will reach approximately half a million triples.

The RDF dataset is largely reuse oriented and directly imports data from the *DBpedia*,[8] *The British National Bibliography*,[9] *MusicBrainz*[10] and `data.gov.uk` datasets. Authority control is provided through cross-checking references between some of these datasets and the *Virtual International Authority File*.[11] The data schema heavily

[5] Inclusion protocols, https://led.kmi.open.ac.uk/node/53/.

[6] The Open University linked data platform, http://data.open.ac.uk.

[7] The accuracy of the above statement over time can be easily verified through a simple SPARQL query on named graphs, as in http://bit.ly/2aiRsm9.

[8] DBpedia, http://dbpedia.org.

[9] British National Bibliography, http://bnb.data.bl.uk/.

[10] MusicBrainz, http://musicbrainz.org.

[11] Virtual International Authority File, http://viaf.org.

relies on external vocabularies such as *Bibo*[12] and the *Music Ontology* [29], as well as providing in-house terminology for listening experiences and additional source document categorisation.

Another by-product of reusing DBpedia data and including LED in a multi-graph linked dataset is the immediate interlinking with learning resources of the Open University, such as courses or study material, that are annotated with reused entities, thus giving way to the possibility of building recommender systems for Open University resources at a minimum overhead.

To accommodate the terms of use under which part of the source manuscripts were licensed to the project, the LED dataset is made available under the *Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License* (CC BY-NC-SA).[13] The project team is investigating on the possibility of partitioning the dataset so that some subsets of the data can be released under more permissive licenses such as CC0.[14]

3.3 Data consumption facilities

Alongside the consumption of the formal data generated by the crowd, which occurs programmatically via standard Linked Data protocols and interfaces (APIs), the fruition of those data by humans is supported by a dedicated implementation of a user interface with RDF support in the back end.

The LED web platform at http://led.kmi.open.ac.uk provides open access to a number of user-centric functionalities for data consumption, including:

- A **tabular data browser** (cf. Fig. 1) designed around a set of dimensions commonly used for exploring data (people, music, source documents) and that integrates external content in the descriptions of its entities;
- An **interactive geographical browser** (cf. Fig. 2) that integrates the above data browser for locations of interest in LED with a Linked Data-powered interactive map;
- A **faceted search** function (cf. Fig. 3) that, on top of a standard keyword search issued by the user and that encompasses both the evidence and the names of all the attached entities, provides a set of filters for restricting search results further. The facets, including the places, people involved, music and sources, are dynamically extracted from the search results as well as from the interlinking of their data with external sources.

As an example of the second facility Fig. 2 shows how the interactive world map can be explored[15] for locations of musical performances or listening experiences in the area of Manhattan, NY. What should be noted here is that, apart from selecting the event location by name (the details of which are given in Sect. 5.1.1), no other information displayed had to be provided by the users in the community at data entry time, nor was it being stored in the dataset to begin with. Data such as alternate names of places, their geographical coordinates, country flags or other representative images, and mereological relationships (e.g. the fact Central Park is located in Manhattan or that Manhattan is part of the USA) were not explicitly provided by the users at any time and were retrieved as necessary from DBpedia and through its linking with GeoNames. Thanks to these labelled relationships in the original data sources, it is often also possible to plot locations of the past such as Prussia or Constantinople, and to relate them with the contemporary political geography.

The rationale described above also governs the plotting of locations of interest on the pages of individual listening experiences (see Fig. 4 for an example).

4 Modelling historical data on listening experiences

There are two orthogonal directions along which a data model for the history of listening to music may be relaxed or constrained, depending on the requirements of the research effort. The union of the *conceptual* domains under consideration—which we shall refer to as the *vertical span* of the research activity—defines the extent to which data should be described depending on the possible interests of users, such as the bibliographical or musical aspect. The *instance* domain—what may be called the *horizontal span*—concerns the characteristics of the settings in which the listening event occurred and were recorded, such as restricting to a specific set of genres, historical period or geographical area. This may imply the restriction of allowed property values to a limited enumerated set, as well as specific classes or properties that only make sense in the given domain. Since LED seeks to provide documented and curated data that conform to a set of scholarly conventions, the vertical span of its model has a natural focus on the bibliographical domain, which is therefore extensively represented. Conversely, LED attempts to strike a balance between the biographical and musical domains, where the variability and richness in information found in the evidence is not guaranteed, for instance due to the possible inability of listeners to precisely recognise or recall the compositions being heard. As its sets out to be the only

[12] The Bibliographic Ontology, http://bibliontology.com.

[13] See http://creativecommons.org/licenses/by-nc-sa/4.0.

[14] CC0 1.0 Universal public domain dedication, https://creativecommons.org/publicdomain/zero/1.0/.

[15] LED geographical browser at https://led.kmi.open.ac.uk/browse/location/.

source

Show 50 entries Filter : war

Title/Description	Author/Editor	Excerpts
Albert Edward Jones 1896–1949 WWI Liverpool Pal Memoirs	Albert E Jones	
BBC WW2 People's War	Ashley Leather, Audrey Lewis, Iain C Macpherson, Joan Styan, John Gardiner, John Kelly, Joseph J Brown, Josie Vernon, Tom Canning	
Court and Private Life in the Time of Queen Charlotte: Being the Journals of Mrs. Papendiek, Assistant Keeper of the Wardrobe and Reader, to Her Majesty – Vol. 1, 2	Mrs Vernon Delves Broughton	4 –
○ Memoir of Charlotte Louise Henrietta Albert Papendiek, Chapter 4, 1778 ○ Memoir of Charlotte Louise Henrietta Albert Papendiek, Chapter 16, 1779? ○ Memoir of Charlotte Louise Henrietta Albert Papendiek, Chapter 22, 1792 ○ Memoir of Charlotte Louise Henrietta Albert Papendiek, Chapter 3, 1776		
Diary of an old contemptible / Edward Roe	Edward Roe, Peter Downham	
Dorothea's War	Richard Crewdson	27 +
Forgotten Voices of the Great War	Max Arthur	
Home fires burning : the Great War diaries of Georgina Lee	Gavin Roynon	13 +
Letter from Anna Seward to Rebecca Cotton Baldwin, 27 October 1785	Anna Seward	
Letters from Headquarters,or Realities of the War in the Crimea – Vol. 1		1 –
○ Letter from Somerset John Gough-Calthorpe, June 1854		
Mud and Bodies: The War Diaries and Letters of Capt. N. A. C. Weir, 1914–1920	Saul David	9 +
My War Diary		5 +
My War Diary 1914–1918		1 +
Parachutes and Petticoats: Welsh women writing on the Second World War	Deirdre Beddoe, Leigh Verrill-Rhys	2 –
○ Second World War reminiscences of Daphne Price ○ Second World War reminiscences of Mair Davies		
Rank and file : the common soldier at peace and war, 1642–1914	T. H. McGuffie	3 –
○ A King's Hussar ○ Extract from The Life of Alexander Alexander. Written by Himself ○ Four Years' Service in India. By a Private Soldier		
Selected Correspondance of Fryderyck Chopin	Arthur Hedley, Bronislaw Edward Sydow (Comp), Chopin	

Fig. 1 Browsing LED by source, highlighting some of the source documents whose parts are modelled in LED as they contain evidence of listening to music

online source of structured data in its category, LED was made to be as lax as possible on the horizontal span, thereby connecting content for any geographical location, genre and period from the 1st century AD onwards, so long as it conforms to the established inclusion protocols and scholarly conventions.

A third dimension, somewhat transversal to those described above, is derived from the crowdsourced nature of certain database systems and concerns the *curation* of data. Informally this means that one may wish to allow only certain types of entities to have their descriptions provided, amended and refined by the community, whilst leaving entities of other types to be simply mentioned without being further described. For example, when indicating that a musical performance took place at the Lyceum Theatre, it may be deemed unnecessary to describe the Lyceum Theatre itself, its geographical coordinates, street address and the fact that it is located in London, England. If that is the case, it is expected that the model will not cater for such features and therefore be limited in its ability to represent the corresponding data. In the remainder of this paper, we will show how this limitation is overcome in LED: the adoption of the Linked Data paradigm and reuse of datasets allows LED to keep the model of *stored* data for non-curated entity types to a minimum, and yet leverage the model of *reused* data of the same entities in order to provide more detailed analytics as well as an improved user experience.

Implementing the data management system of LED natively in RDF, the basic format of Linked Data, means that the model described here and used for the data that are

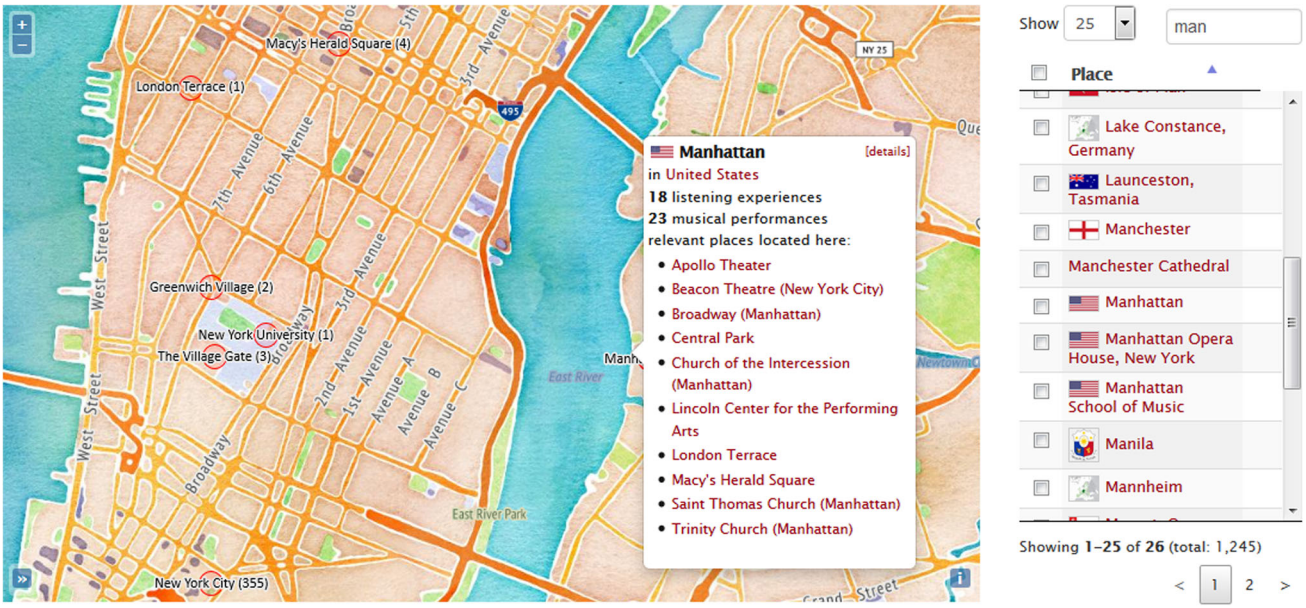

Fig. 2 Geographical browser highlighting places in or near Manhattan that are of interest for LED data; map rendered in OpenLayers 3 (http://openlayers.org) using the Stamen watercolour tile server (http://maps.stamen.com)

exposed to the public is the same as the one that is used to store them internally, barring a few unnecessary metadata attributes. The machine-readable version of the LED ontology in the OWL language [36], inclusive of all its external dependencies, is available at https://led.kmi.open.ac.uk/ontology.

4.1 Model of curated data

This section describes the ontological model of the types of digital objects that the users are explicitly allowed to describe as part of their input macrotasks. In the following, we will assume any used prefixes to correspond to specific data vocabularies as in Table 1.

4.1.1 Listening experiences

The LED data management workflow revolves around the **Listening experience** object class: the macrotask of submitting a new listening experience is that through which all the other digital objects in the curated part of the model are described. A listening experience (LE) is a documented engagement of an individual in a performance event of one or more pieces of music, where "documented" means that there must be a quotable and traceable source, referring to which is mandatory.

Figure 5 shows the representational schema for a LE as Linked Data. Each node in the graph denotes a class of items that can be related with one another via a property represented by an arc going from subject to object. Every arc is labelled with the corresponding RDF property names. Placeholders for the detailed models described in the next sections are also included.

It is shown that every `led:ListeningExperience` is also an `event:Event`, reflecting the rationale by which listening experiences are *subjective events* that take place within the innermost context or each listener, especially the one reporting the evidence. In so doing, the flexibility of the *Event Ontology*,[16] one of the most common ways of representing general events in the Semantic Web, is also inherited. Properties from the Event Ontology such as `event:agent`, `event:place` and `event:time` are utilised and put into effect in order to denote participants (particularly listeners), locations and instants or intervals when the subjective event took place. The connection to the music being heard can be of one of two kinds, i.e. either by referencing the musical work directly, if there is nothing to be said about the context of its performance, or by describing the performance itself as prescribed according to Sect. 4.1.3. In the latter case, all the listeners (who are typed as `foaf:Agents`) are also linked to the performance itself through the `mo:listener` property of the Music Ontology.

4.1.2 Sources and excerpts

Describing the documents where the written evidence of listening was found is by far the most extensively modelled

[16] Event Ontology specification, http://motools.sourceforge.net/event/event.html.

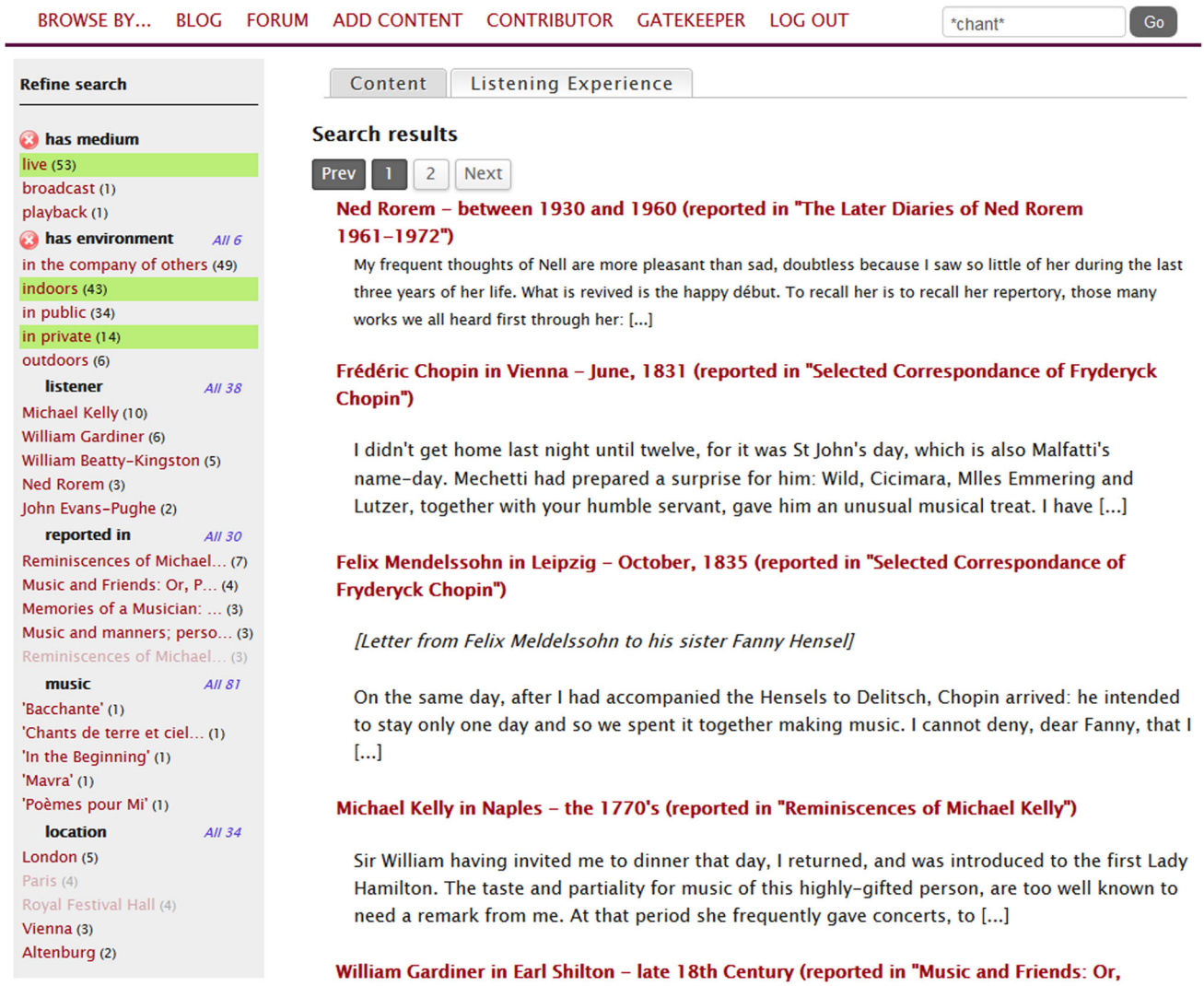

Fig. 3 Example of faceted search on the "chant" keyword, restricting search results to private indoor live performances

aspect of LED, which was for its vast majority served by the *Bibo* ontology, arguably the most widely utilised vocabulary of linked library data. Although the only mandatory attributes of a source in LED are its type (e.g. `bibo:Periodical`, `bibo:Letter` or `bibo:Book`), name and publishing status, much more detailed information can be provided if available. Figure 6 shows the portion of interest of Bibo that is used for LED, along with its dependencies including the Dublin Core metadata terms.[17] It is possible to indicate the set of authors, original language, translators into English and any Web reference of note. Specifically for published sources, it is also possible to provide applicable data such as the editors, publishing organisations, ISBN (10- or 13-digit), publication time and place, volume number and issue number (if an academic or news article).

In order to appropriately model enhancements to the bibliographical data that are re-published by LED, the hierarchical organisation of sources into a *meronomy* is formally represented. Should the inputter choose to describe a specific portion of a published book or collection, such as a chapter, article, letter or diary entry, the LED model has the flexibility to allow authorship data to be filled in for that portion, whilst being linked to the original container document via a `dc:isPartOf` relation. A further level in the hierarchy is occupied by the text excerpts themselves, i.e. the quotes from the source that contain the detailed evidence of a listening experience. These are modelled as instances of `bibo:Excerpt` and linked to the source of either of the above levels via a further `dc:isPartOf` relation, with an optional indication of the page numbers where it appears. The actual text content is encoded as a value for the `rdf:value` property, though we have recently started allowing formatted text, which is encoded in HTML and entered through a rich

[17] Dublin Core Metadata Terms, http://dublincore.org/documents/dcmi-terms/.

Fig. 4 A listening experience rendered on the LED portal

text editor on the input user interface. Although an excerpt may have multiple versions of the same text with different HTML formatting, the uniqueness and invariability of the URI that identifies the excerpt is guaranteed by computing the identifier from the plain text variant (i.e. stripped of any HTML tags) of its entire `rdf:value` and the URI of the source document. Therefore, the URI only changes if the actual text content does.

4.1.3 Music

The information concerning what was actually being heard by the listeners in the recorded evidence can display the highest variability, compared to the other dimensions whose studies LED supports: due to the wide social and cultural spectrum of potential listeners, the degree of knowledgeability in the musical domain is highly mutable, which may lead listeners to refer to the music being heard by given title, or by genre, or by author, etc., and provide descriptions equally variable in richness. On these assumptions, later confirmed by the submitted data, "Music" is here intended as an umbrella term that captures musical performances as well as the immaterial musical works upon which they are based, and that may be given without further details on their performances. This is reflected in the ontological model of music as per Fig. 7, where `mo:MusicalWork` and `mo:Performance` are the classes of entities that may directly be linked to a `led:ListeningExperience`.

LED's focus on the ontological model of music is on those features of a musical work or performance that pertain to the

Table 1 Prefix-vocabulary mappings for the LED schema

Prefix	Vocabulary	Namespace
bibo	Bibliographic Ontology	http://purl.org/ontology/bibo/
dbpo	DBpedia Ontology	http://dbpedia.org/ontology/
dc	Dublin Core metadata terms	http://purl.org/dc/terms/
edtf	Extended Date-Time Format (custom)	http://data.open.ac.uk/edtf/ontology/
event	Event Ontology	http://purl.org/NET/c4dm/event.owl#
foaf	Friend-Of-A-Friend	http://xmlns.com/foaf/0.1/
geo	WGS84 Geo Positioning	http://www.w3.org/2003/01/geo/wgs84_pos#
gs	GeoSPARQL	http://www.opengis.net/ont/geosparql#
led	LED internal vocabulary	http://led.kmi.open.ac.uk/term/
mo	Music Ontology	http://purl.org/ontology/mo/
org	W3C Organisation Ontology	http://www.w3.org/ns/org#
owl	Web Ontology Language terms	http://www.w3.org/2002/07/owl#
rdf	Standard RDF vocabulary	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	RDF Schema vocabulary	http://www.w3.org/2000/01/rdf-schema#
schema	Schema.org	http://schema.org/
swat	SWAT Project upper ontology	http://www.co-ode.org/roberts/upper.owl#
time	OWL-Time	http://www.w3.org/2006/time#
travel	SWAT Project ontology of travelling	http://www.co-ode.org/roberts/travel.owl#

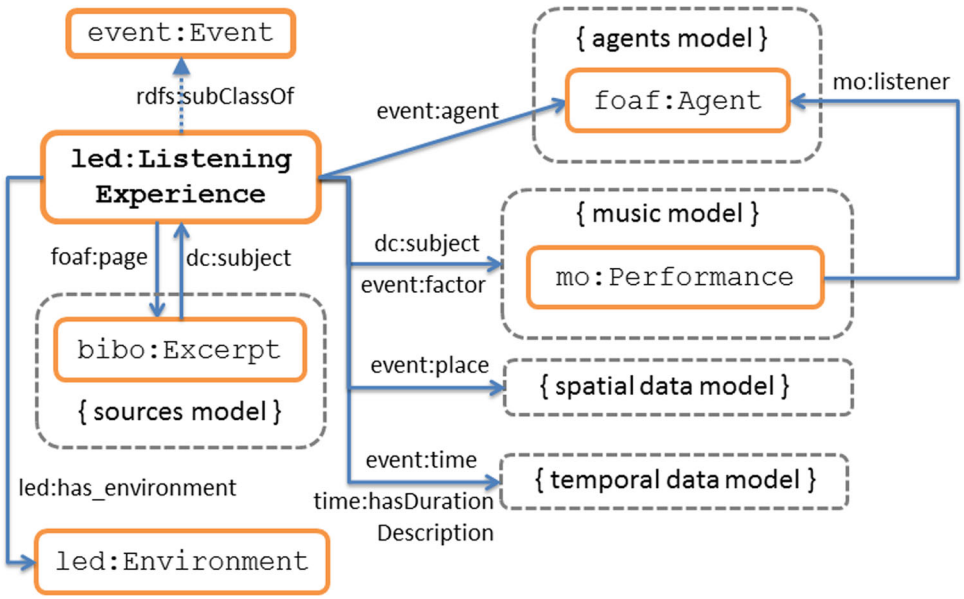

Fig. 5 Ontological model of listening experiences

listener's perception of the events themselves, namely the identities of the performers, the genre of the performance, instruments used, location and other environmental information (such as whether the performance occurred indoors or in a public space). Data originating from any prior knowledge the listener may possess concerning the music being heard are not contemplated, apart from the identities of composers and the intrinsic ability of a listener or scholar to refer to a work, musician, genre or instrument by name. Likewise, the material realisation of musical works, such as record albums or MP3 files, was not expected to have a sufficient impact on the perception of music as to warrant the representation of the discographical domain (e.g. release dates, production staff or record companies); therefore, this element is not present in the LED data model. The vocabulary used is almost entirely contained in the *Music Ontology* specification, an open standard adopted by several data providers, including the BBC and Linked Data versions of MusicBrainz.

In accordance with what the Music Ontology specification prescribes, LED is relatively lenient with regard to

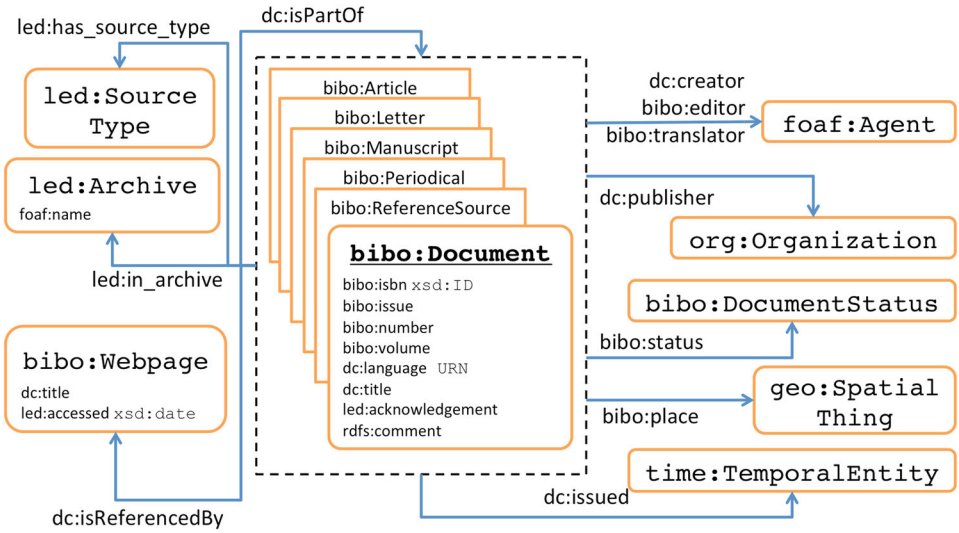

Fig. 6 Ontological model of sources

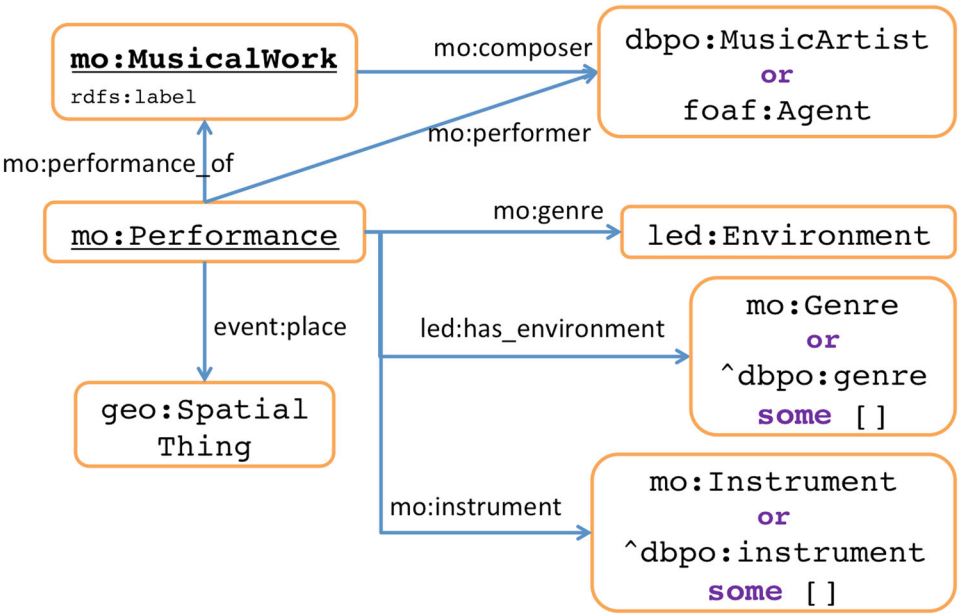

Fig. 7 Ontological model of musical entities

the taxonomies used for the values of `mo:Genre` and `mo:Instrument`. In fact, during the first two years of project run-time, users were allowed to enter arbitrary values for both fields, to retrospectively align with the listener's perception of the types of music and instruments at the time. After the experimental run, we compiled the sets of input values and attempted to align them with taxonomies such as the DBpedia ones, extending the scope of the concepts accordingly. Consequently, the notion of genre with respect to its admissible values has since been extended to include not only musical genres per se, but also compositional forms (e.g. "concerto", "solo") and song forms (e.g. "chanson", "anthem"). Likewise, the notion of instrument covers several instance levels, ranging for instance from "strings" to "guitar" and all the way down to "Gibson Les Paul"; it was also extended to include anything that is known to have been used as an instrument (e.g. telephones or 8-bit computers) as well as musical techniques (e.g. "double drumming", "stride") and vocal styles (e.g. "tenor", "coloratura"). Some of the heuristics used for detecting the candidate extensions to the vocabularies for genres and instruments operated through lookup of other datasets linked by DBpedia, such as *Freebase*.[18] Since the end of this experimental phase, new values can no longer be proposed by the users; however, no requests to add missing values have since been reported by the community.

[18] Freebase, https://developers.google.com/freebase/ (now discontinued for Google's Knowledge Graph API).

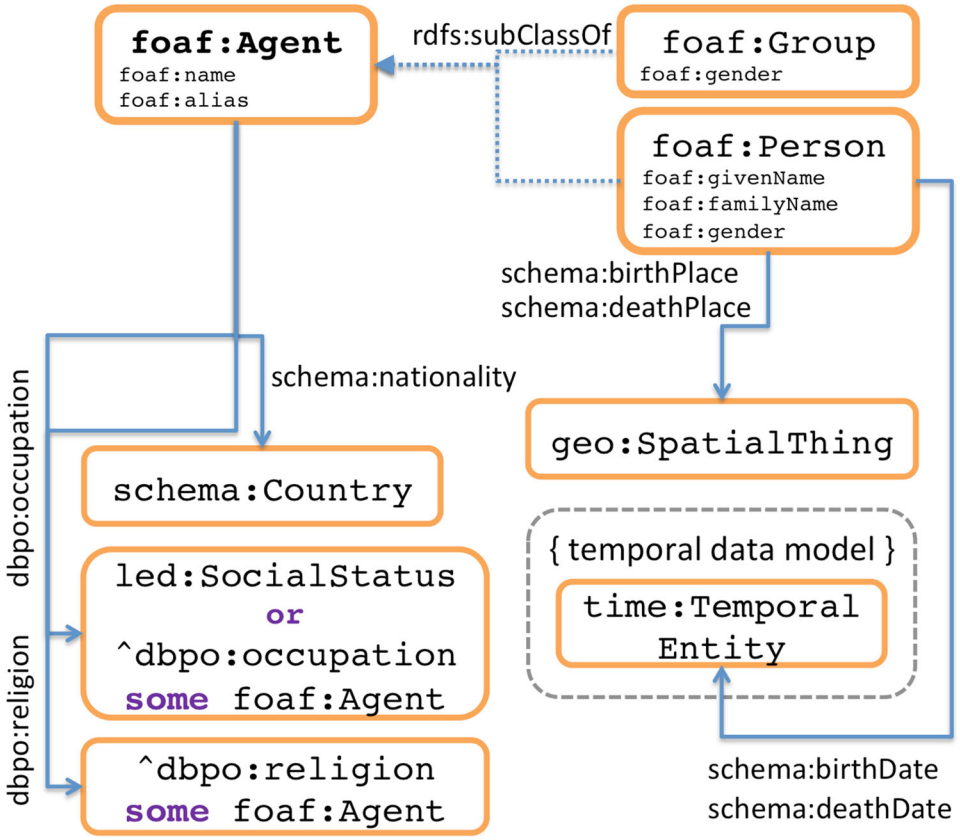

Fig. 8 Ontological model of agents

4.1.4 Agents

`foaf:Agent` is the aggregating general class for all individuals in LED that serve a role as listeners, composers, performers, writers, editors or translators at any point in the dataset. They may, if this information is given, specialise to a single person or a group (e.g. a band or audience). The model or data captured in LED (cf. Fig. 8) uses a combination of the *FOAF* ontology and the *Schema.org* general-purpose vocabulary for Web annotation, with only a few residual properties lifted from the DBpedia ontology.

To facilitate search and discovery by multiple labels, either birth names or alternate names such as honorary titles or pen names are supported. In addition to basic anagraphical information (birth/death places and dates, nationalities and genders, possibly mixed for groups), data extracted from the social and cultural context of the agents are also accommodated, such as religions and occupations. The admissible values for both these properties were initially extracted through reverse lookup of their usage throughout DBpedia, which has proven sufficient for religions or philosophies of life, as all the prior arbitrary values entered by the users could be effectively aligned. As for occupations and, more in general, indicators of the agent's position in the social ladder, the vocabulary was extended using a combination of the ISCO08 taxonomy[19] and the National Readership Survey demographic classification,[20] both made available for the first time in RDF.

4.2 Model of non-curated data

This section describes the ontological model of the types of entities that are not directly described by the users, namely those that encode spatial and temporal information.

4.2.1 Spatial information

The model of spatial data in LED underwent a major overhaul during the second half of the project, moving to sheer reuse of exactly named geographical locations to a flexible but controlled structure. In the updated model, it is possible for contributors to specify whether an experience took place on a journey between two locations, as well as to enter arbitrary content quoting the listener's own denomination of where it took place, without losing the ability to reuse named geographical locations. Whilst a detailed description of the

[19] International Standard Classification of Occupations, http://www.ilo.org/public/english/bureau/stat/isco/isco08/.

[20] National Readership Survey social grade, http://www.nrs.co.uk/nrs-print/lifestyle-and-classification-data/social-grade/.

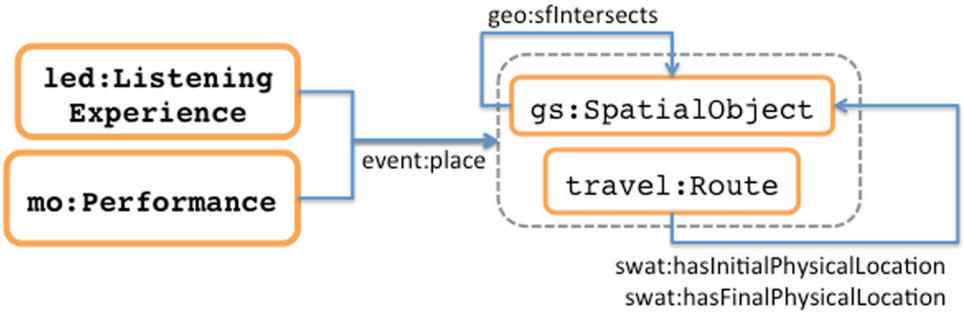

Fig. 9 Ontological model of spatial data in LED

techniques used to manage the input according to this model will be given in Sect. 5.1.1, the model itself is summarised in Fig. 9.

A listening experience or musical performance can be linked, via the `place` property of the Event Ontology, either to a defined spatial object, which may in turn intersect with other spatial objects (e.g. "in church in Birkenhead" and "a synagogue in Paris" intersect with `dbpedia:Birkenhead` and `dbpedia:Paris`, respectively), or to a route with an initial and a final location. As existing ontologies to represent journeys are scarce, one of the experimental ontologies produced by the SWAT project[21] was used, as it offered the desired trade-off with the simplicity and flexibility requirements deriving from the early user experience of contributors.

As for countering the vagueness and consequent proliferation of URIs deriving from inputting arbitrary text for locations, it was a design decision to use entities that denote *sets of locations* alongside exact locations. When a precise named entity is entered, regardless of the granularity, it is directly reused; if on the other hand part of the input text cannot be linked to a place, the database will use the one entity that implicitly denotes all the places that share the given description. For instance, all the listening experiences that occurred in a "village in Sierra Leone" will point to the same URIs that is the generic representation of any village in Sierra Leone, until a revision or refinement established that one listening experience occurred in a specific village.

4.2.2 Temporal information

Great attention was also devoted to providing flexibility in the modelling of temporal data, relating especially to the time when a listening event took place, but also when providing birth and death dates of listeners, publication times, etc. Peculiar to this notion is the need for a structured representation in RDF of data that are often partial, inaccurate or fall within an approximate interval. There are no set standards for encoding expressions such as "the 1920s" or "sometime in June in the 20th Century" as standard literal values, which rules out the possibility of adopting standard XML Schema date and time literals throughout the dataset. Alongside establishing a convention for encoding these formulae, it was also a requirement that the resulting RDF should easily respond to simple SPARQL queries that take partial event data into account.

Figure 10 shows how temporal data relating to listening experiences are modelled. The core classes for intervals, durations and temporal entities in general are lifted from the OWL-Time vocabulary [13]: it is noteworthy that `time:Interval` is here used to indicate a confidence interval wherein the event may have happened and not its entire duration, which is instead deferred to the `time:hasDurationDescription` property.

To represent vague or underspecified temporal entities, the EDTF draft proposal of the Library of Congress [24] was employed. EDTF provides a syntax for representing partial dates, decades, seasons, approximations, etc., in a similar fashion as XML Schema literals: traces of its adoption in MARC formats were found in the Yale University Music Library[22] and other online catalogues. In LED, EDTF values are represented as URIs and fully modelled in RDF, with a few extensions to represent subjective fuzzy intervals like "early" or "late"; this is, to the best of our knowledge, the first attempt at providing a Linked Data version of EDTF. We acknowledge parallel efforts in representing time as Linked Data such as the PeriodO gazetteer [16]; however, the goal of LED was to provide a uniform convention rather than reuse historical denominations of specific periods.

5 Data management workflow

Implementing a supervised crowdsourcing governance model as per Sect. 3.1.1 in a purely Linked Data system was an unprecedented effort at the time LED was established. Not only do the typical states of contributed resources and their transitions need to be realised in terms of their impact on an

[21] Semantic Web Authoring Tool, http://mcs.open.ac.uk/nlg/SWAT/.

[22] Music cataloging of Yale: MARC 046 tagging, http://web.library.yale.edu/cataloging/music/MARC046.

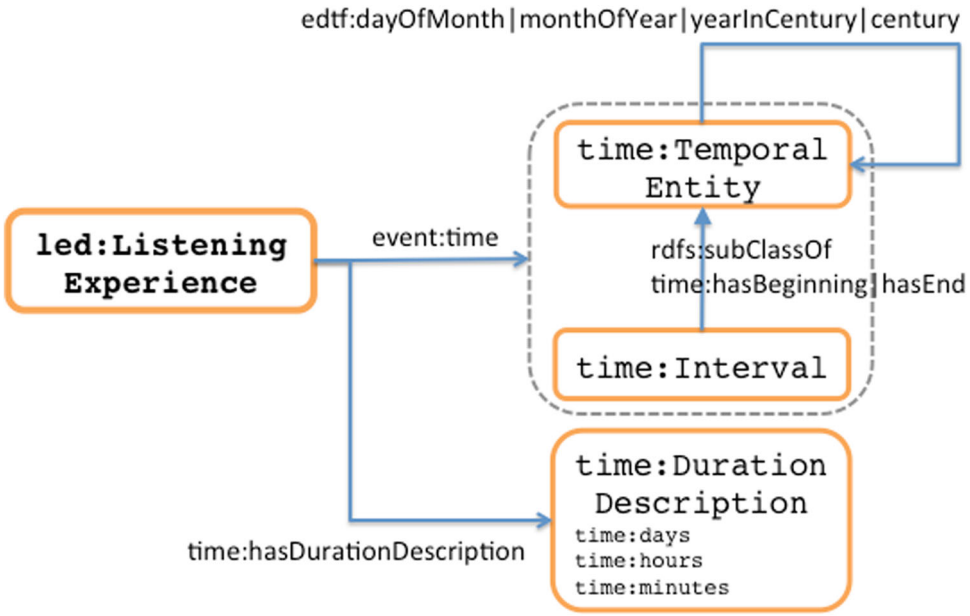

Fig. 10 Ontological model of temporal data for listening experiences

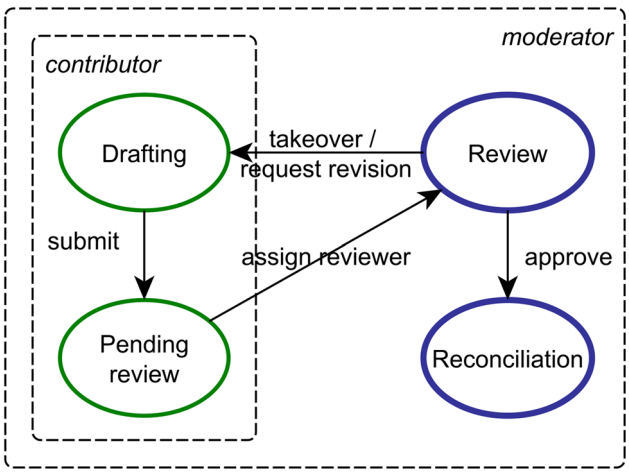

Fig. 11 Management workflow of entity data

RDF storage system, but also further operations need to be considered in the light of the strong identity of Web resources that is implied by the Linked Data paradigm [18].

As illustrated in Fig. 11, the resources that can be described in LED, i.e. listening experiences and the entities that play a role as their factors, undergo a data management treatment that can be summarised as being structured into four phases:

- During **drafting** the data being originally authored by the contributor, thus excluding those being reused, are only visible to the contributor themselves. The draft state is the initial state for any data entry process.
- A resource is **pending review** when the contributor has completed their draft and submits it to the attention of moderators. Its data then become visible to all moderators, as well as to the contributor, who however cannot reclaim the submission or propose amendments. Any data that add to or modify approved data are considered as claims made by the user, which will not propagate until approved.
- The **review** phase occurs when a moderator is assigned to a listening experience pending review and either promotes it to public, takes it over for editing, or sends it back to the contributor for revision.
- **Reconciliation** applies to resources promoted to public visibility that semantically denote the same entity, and results in either merging their data or establishing mappings between them.

Whilst the phases of drafting, pending review and review phases are typical of most supervised data sourcing frameworks, reconciliation depends on the notion of identity that is adopted; as this is very strong in Linked Data thanks to the adoption of URIs and alignment properties in the representation languages, it was included in the data management workflow of LED.

5.1 Linked Data reuse and enhancement

The philosophy behind data reuse in LED is to drive the data management workflow so that any facts generated by the crowdsourcing community directly reference the entities they reused. In Linked Data terms, this means that the object of a generated RDF triple is, whenever possible, the external URI itself: in general, LED does not create new URIs for reusable entities to align with at a later stage, though this is possible if one so desires. Rather, LED aims at being an

additional trustworthy source of exclusive information about entities that are already well known in the Linked Data cloud. The datasets reused by LED, their extent and the types of enhancements provided on top of each are detailed below.

British National Bibliography (BNB) has a Linked Open Data counterpart made available directly by the British Library.[23] This dataset records the publishing activity of the UK and Ireland since 1950, including periodicals and collections. It also provides basic biographical information about authors and contributors, such as birth and death dates. Besides information about authors and editors (conveyed, however, through generic Dublin Core metadata properties), the span of BNB data is mostly the publication information of its entries. LED publishes refinements of these data that take into account the specific roles of contributors (authors, editors or translators from foreign languages), external Web references and up to two additional mereological levels for each published item, i.e. specific documents contained therein (e.g. letters, journal entries, articles) and the very same text excerpts that constitute the core evidence of listening experiences.

DBpedia is a general-purpose dataset obtained by refactoring the structured content found in Wikipedia pages, such as infoboxes.[24] It is crowdsourced by a very vast community, therefore the levels of detail and accuracy tend to vary greatly, however they are moderately harmonised and cleaned up by the DBpedia treatment. DBpedia is the primary source of data reuse in LED and, with possibly the sole exception of published sources, its URIs tend to be preferred to those from other datasets on the same entities, due to their high popularity and reuse rate in the Linked Data cloud. The vocabularies used by LED for instruments, musical genres and religions are also lifted from DBpedia. Any missing information that is contributed by the LED community over DBpedia entities is re-published, thus enriching the information provided by the source dataset: the most frequent enhancements found in LED are census data of listeners, namely their occupations, nationalities, social status and religions.

MusicBrainz is a public domain database of encyclopedic content that aggregates data about musicians, musical works and their relationships.[25] Although direct support for RDF has since been discontinued by MusicBrainz itself, the permissive license on the data has allowed third parties to maintain LD counterparts. LED references artists, musical works and recordings dating as of the latest 2015 data dump provided by the LinkedBrainz project.[26] Re-published enhancements include specifications of performances of musical works in MusicBrainz, relations to composers, biographical data for artists and alignments to BNB and DBpedia.

VIAF is one of the largest online resources for authority control.[27] Though it does not currently provide a SPARQL endpoint, every VIAF item can be exported to RDF. VIAF entities are not directly reused in LED; however, their links to DBpedia and BNB URIs are used in order to collapse equivalent entities as one and the same and reduce the likelihood of duplicate recommendations in the data entry process.

data.gov.uk is a massive hub of British eGovernment datasets. When an exact, fully qualified date is entered in LED, it is converted to an entity by its own right that references the British calendar specification provided by *data.gov.uk*,[28] rather than a simple literal. This allows us to use similar RDF descriptions for exact dates and partial time specifications as in our custom RDF version of the EDTF specification.

5.1.1 Extraction of Linked Data from natural language input

Whilst conscious data reuse has proven effective for most constituent elements of a listening experience in the LED model, it was perceived as overly restrictive in the definition of the geographical locations where the experiences were recorded. It was often the case that sources would reference a location that is not an exact match with a place of public interest (e.g. of the form "at X's house in Y"), and as such does not appear in geographical datasets.

From a purely record-keeping point of view, one solution would be to relax data entry for locations so that a contributor may freely enter arbitrary text. From a Linked Data perspective, however, this has an inherent risk of generating unlinkable data, which would defeat the purpose of adopting this paradigm in the first place. One approach involves mining the text with data-aware natural language processing (NLP) engines, but the need to store their output in RDF raises two challenges: (i) how to prevent multiple entries of the same text from generating redundant URIs; and (ii) which properties should be used to link the places extracted from the text with one another and with the place represented by the text. This prompted a hybrid approach recently implemented in LED, where *supervised* named entity extraction is performed on the text as the user types their input.

Figure 12 shows an example of user interaction with supervised named entity extraction on the location field. Here,

[23] British National Bibliography, http://bnb.data.bl.uk.

[24] DBpedia, http://dbpedia.org.

[25] MusicBrainz, http://musicbrainz.org.

[26] LinkedBrainz, http://linkedbrainz.org.

[27] Virtual International Authority File, http://viaf.org.

[28] See, e.g. http://www.epimorphics.com/web/wiki/using-interval-set-uris-statistical-data.

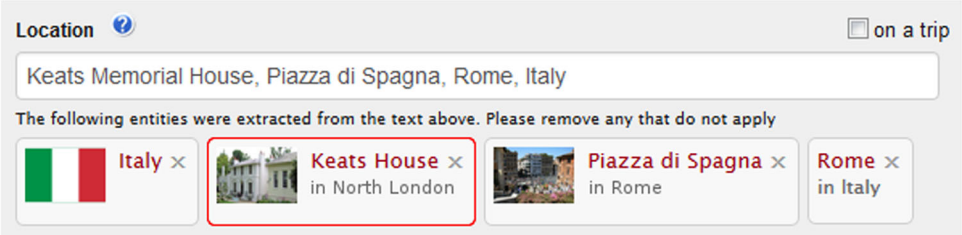

Fig. 12 Example of supervised named entity extraction from places. Here "Keats House" would be removed by the user because the entity displayed—i.e. the Keats House in Hampstead, London—does not correspond to the place the user intended—i.e. the Keats–Shelley Memorial House in Rome

a specially zealous user types "John Keats' House Piazza di Spagna, Rome, Italy" in an attempt to avoid confusion with John Keats' house in Hampstead, London. The underlying NLP engine—more precisely a pipeline of *OpenNLP*[29] processors for language detection, sentence recognition and part-of-speech tagging, in turn post-processed by a bespoke *Stanbol* enhancement engine[30] for linking with places in DBpedia—extracts four distinct places from the input text. However, as one of the recommendations is Keats' house in London due to being an exact syntactic match (the one in Rome being actually called Keats–Shelley Memorial House), the user will be able to reject that recommendation.

The RDF triples generated by the process are as follows: a single triple is generated with the listening experience as subject and `event:place` as predicate; if only one place was extracted from the text and its name is an exact syntactic match, the URI of that place is the object; otherwise, the object URI is calculated by hashing the list of recommended URIs sorted alphabetically and the (case-insenstive) input text: this guarantees that repetitions of the process reuse the same URIs as needed. The generated URI is linked to the recommended places extracted from DBpedia by materialising triples over the `geosparql:sfIntersects` predicate: this is due to the fact that, in most cases, the combination of several places entered together can be interpreted as a topological intersection (i.e. that the event occurred, not necessarily simultaneously, "somewhere in all of these places"). The only likely exception to this rule is proximity (e.g. "near Naples"), but only four such occurrences have been detected so far.

This solution has allowed us to keep the number of generated URIs under control, reduce the need for data reconciliation and guarantee a machine-readable semantic interpretation of the geographic input that closely approximates the one intended by inputters. Future improvements currently under consideration include the interpretation of topological keywords (e.g. "near", "in") and their translation into corresponding GeoSPARQL statements, suggestion of alternatives in the supervised phase and support for entity linking with GeoNames.

5.2 Data shaping and materialisation

Implementing a human database interface to work with a native Linked Data system means ensuring that the underlying RDF store is able to provide real time responses to all the SPARQL queries that encode the ways to consume the data through the user interface. Best practices such as the Linked Data API specification for resource consumption[31] are a valuable aid to the optimisation of data and queries, however the multidisciplinary domain of music listening experiences in the literature called for additional facilities for appropriately consuming LED data (cf. Sect. 3.3) beyond the scope of these specifications.

One example where additional optimisation efforts are warranted is by browsing LED "by people",[32] which entails individuals, music ensembles, groups in general and persons with different and multiple roles. A single SPARQL query that satisfies all the corresponding RDF types and properties that denote each role showed to take several times as long to answer as if the same entities satisfied the constraint of a single RDF type.

In the spirit of these examples and consideration, and taking into account the studies carried out on the impact of SPARQL querying on datasets that contain materialised (redundant) inferences [2], the following adjustments were implemented in the storage mechanisms of LED data:

1. At least one generic value for `rdf:type` is materialised into the RDF store for every entity of a type that LED can be browsed for: in the case of persons, groups, ensembles, etc., the type is `foaf:Agent`.
2. The selected `rdfs:label` of any entity that is reused at data entry time is stored in the RDF dataset, even though the user may not be allowed to curate it. This allows tabular views to be sortable via SPARQL and search indices

[29] Apache OpenNLP, http://opennlp.apache.org.

[30] Apache Stanbol, http://stanbol.apache.org.

[31] Linked Data API specification, https://github.com/UKGovLD/linked-data-api.

[32] Browse by people, https://led.kmi.open.ac.uk/browse/people/.

to be created and managed by the triple store for full-text searches.
3. No blank nodes are ever created.

5.3 Reconciliation

Data reuse in LED is a conscious process, in that it only occurs if the contributor intentionally selects the entity to be referenced as the listening experience record is created [10]. In order to tackle homonymy issues and guarantee the coexistence of multiple entities with the same name, the strategy is to generate an entirely new URI for any entity for which the user inputs data but does not reuse an existing entity as a subject. To deal with the event that users may overlook the data reuse capabilities of the system, thus generating redundant URIs for the same entity, a data reconciliation feature was introduced after the database reached a critical mass of 500 records. Through this functionality it is possible to select any number of similarly named entities of the same category, designate the one whose identifying URI is to be preserved (the *primary* entity) and decide whether the data entered for the entities to be aligned (*secondary* entities) should be merged with those of the primary one or discarded altogether [10].

Reconciling LED affects the underlying RDF dataset as follows. All the triples having secondary entities as objects are always rewritten with the primary entity as object. The signatures of secondary entities (i.e. the triples having them as subjects, as well as those that have no other links than the secondary entities) are deleted and only selectively rewritten for the primary entity if the user has opted to do so. Additionally, if any secondary entity is a reused external one (e.g. from BNB or MusicBrainz), an `owl:sameAs` link is generated between that and the primary entity. This has the effect of producing an additional enhancement on top of existing linked data, in the form of alignments between datasets that are only sparsely, if at all, aligned; for instance, only a small subset of LinkedBrainz is aligned with DBpedia, and the BNB dataset is not directly aligned with DBpedia, if not through the VIAF dataset.

Data reconciliation is currently a privileged feature available only to LED moderators. The possibility of opening it to the community, thus supporting crowdsourced data linking as well, has been investigated, but will not be implemented as long as the project is able to commit resources from its own team. Closer to realisation is the extension of this feature to support non-named entities, such as musical performances and listening experiences.

6 Discussion

The Listening Experience Database is used in teaching on a number of courses in music at The Open University, such as *A342 Central questions in the study of music*,[33] in which students are asked to use it to research listeners' reactions to recorded music; and *A873 MA in Music*,[34] in which it forms the subject of a digital humanities case study. It is also used in the *Critical Portfolio* module, which forms part of the Master of Music (MMus) in Performance at the Royal College of Music.[35]

As a testament to its external impact, the content of LED has been the subject of a recent study on the sensory impact of listening [20]. In addition, the database is associated with the BBC World Service radio series *The Music of Time* (2017)[36] and is cited at the end of each episode with an invitation to listeners to explore and contribute to it. It was used by the production team in the pre-reproduction stages of the program to familiarise themselves with the aims of the project.

Whilst the topic mapping of the dataset is one trail of ongoing investigation for the project, empirical analyses of the data have shown them to be skewed towards specific genres, historical periods and geographical areas. Most notably there is a predominance of accounts of listening to various forms of art music (e.g. classical, operatic or sacred) in nineteenth- and twentieth-century Britain. This concentration reflects the research interest of the team that originally established LED [9,19,28] and provided material for further studies by other researchers in music history [25,39]. The fact that the database at present accepts only entries in the English language tends to contribute to the geographical bias, as does the nature of previous efforts targeting the availability of textual sources on music listening. Examples of the latter include the *Calendar of London Concerts 1750–1800* database,[37] the *19th-Century London Concert Life 1815–1895* project[38] and the collections of concert programmes held in European libraries, archives and museums.[39]

Although the biases from the historical context are not expected to change in the near future, given that the focus of the second phase of the project is specifically on listening in Britain since 1700, steps are being taken in order to highlight gaps and address them as far as is practicable. For example, the paucity of early (pre-1800) sources is currently being addressed internally by the project team, whereas the design

[33] Central Questions in the study of music, http://www.open.ac.uk/courses/modules/a342.

[34] MA Music part 1, http://www.open.ac.uk/postgraduate/modules/a873.

[35] Masters (MPerf, MComp, MMus) Programme at a glance, http://www.rcm.ac.uk/media/RCM%20Masters%20Programme%20at-a-glance.pdf.

[36] The Music of Time http://www.open.edu/openlearn/tv-radio-events/radio/the-music-time.

[37] See http://research.gold.ac.uk/10342/.

[38] See http://www.concertlifeproject.com/index.html.

[39] See http://www.concertprogrammes.org.uk.

of analytics tools and aggregated views on the data is also targeting this line of investigation.

The project is currently funded until mid 2019, during which time software development, data inputting and curation, dissemination and public engagement activities will be covered. Through its academic network LED has engaged assistants outside the project team who are responsible for checking and uploading outstanding submissions, clearing the remaining 10% backlog in the process. The Open University is committed to maintaining the accessibility of the LED Web portal and content for three years from the end of the funded project, in accordance with the requirements of Research Councils UK.[40] The sustainability of the machine-readable data is also guaranteed by their mirroring on the *data.open.ac.uk* platform, which is a separate project fully supported by the OU. The curation of selected attributes of reused entities from external datasets guarantees the functionality of the LED dataset as a stand-alone resource, even in the event that the external sources become unavailable or change drastically, as was already demonstrated during a prolonged LinkedBrainz downtime.

7 Conclusions

This article has described LED, a crowdsourced linked dataset that aggregates exclusive data about experiences of listening to music throughout history. The LED dataset reuses and publishes enhancements of linked data from bibliographical and musical sources, as well as from general-purpose sources. To our knowledge, this is not only the first research effort in gathering and assembling evidence of listening as found in the literature and social media alike, but also the first attempt at crowdsourcing a linked dataset whose entire life cycle is implemented using the technologies of its foundational standard.

Over eighteen months since being introduced to the digital library community [10], the dataset has grown tenfold and new features have been added to both the dataset and the portal: these include support for data from MusicBrainz and *data.gov.uk*, interactive geographical browsing, a convention for representing vague temporal data and named entity recognition for arbitrary location text.

Support for crowdsourcing microtasks, partitioning the dataset to support permissive licensing and possible synergies with the DOREMUS project output, have already been mentioned as ongoing investigation into the evolution of LED. Further work planned for the current iteration of the project includes the addition of tools for providing users with analytics on top of the richness of LED data as well as by directly mining the evidence text, after a first successful run of experimental text mining on location descriptions. Particularly, there is interest in detecting implicit communities and clustering entities in LED based on dimensions of interest such as cultural and sociological aspects. Another direction in which LED will evolve is the growth of its dataset based not only on the already implemented data entry process, but also on more direct input methods, such as auto-filling an entry after bookmarking a corresponding web page that reports evidence of listening. Implementing Web crawlers to detect and aggregate listening experiences from unexplored sources on the Web is another task currently underway.

[40] Research Councils UK, http://www.rcuk.ac.uk.

References

1. Achichi, M., Bailly, R., Cecconi, C., Destandau, M., Todorov, K., Troncy, R.: DOREMUS: doing reusable musical data. In: S. Villata, J.Z. Pan, M. Dragoni (eds.) Proceedings of the ISWC 2015 Posters & Demonstrations Track Co-located with the 14th International Semantic Web Conference (ISWC-2015), Bethlehem, PA, USA, October 11, 2015, *CEUR Workshop Proceedings*, vol. 1486. CEUR-WS.org. http://ceur-ws.org/Vol-1486/paper_75.pdf (2015)
2. Ahmeti, A., Polleres, A.: SPARQL update under RDFS entailment in fully materialized and redundancy-free triple stores. In: I. Celino, E.D. Valle, M. Krötzsch, S. Schlobach (eds.) Proceedings of the 2nd International Workshop on Ordering and Reasoning, OrdRing 2013, Co-located with the 12th International Semantic Web Conference (ISWC 2013), Sydney, Australia, October 22nd, 2013, *CEUR Workshop Proceedings*, vol. 1059, pp. 21–32. CEUR-WS.org. http://ceur-ws.org/Vol-1059/ordring2013-paper3.pdf (2013)
3. Barlow, H., Rowland, D. (eds.): Listening to music: people, practices and experiences. http://ledbooks.org/proceedings2017/ (2017)
4. Basharat, A., Arpinar, I.B., Dastgheib, S., Kursuncu, U., Kochut, K.J., Dogdu, E.: Semantically enriched task and workflow automation in crowdsourcing for linked data management. Int. J. Semant. Comput. **8**(4), 415–440 (2014). https://doi.org/10.1142/S1793351X14400133
5. Bekiari, C., Doerr, M., LeBoeuf, P.: FRBR: object-oriented definition and mapping to FRBR-ER. Technical Report, International Working Group on FRBR and CIDOC CRM Harmonisation. http://www.cidoc-crm.org/docs/frbr_oo/frbr_docs/FRBRoo_V1.0_2009_june_.pdf (2009)
6. Bertin-Mahieux, T., Ellis, D.P.W., Whitman, B., Lamere, P.: The million song dataset. In: A. Klapuri, C. Leider (eds.) Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011, Miami, Florida, USA, October 24–28, 2011, pp. 591–596. University of Miami. http://ismir2011.ismir.net/papers/OS6-1.pdf (2011)
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data–the story so far. Int. J. Semant. Web Inf. Syst. **5**(3), 1–22 (2009)
8. Bradley, M.: The reading experience database. J. Vic. Cult. **15**(1), 151–153 (2010). https://doi.org/10.1080/13555501003607792

9. Brown, S.: Analysing listening experiences: a case study of the young Benjamin Britten. In: Barlow and Rowland [3]. http://ledbooks.org/proceedings2017/#sec_210_h1
10. Brown, S., Adamou, A., Barlow, H., d'Aquin, M.: Building listening experience linked data through crowd-sourcing and reuse of library data. In: B. Fields, K.R. Page (eds.) Proceedings of the 1st International Workshop on Digital Libraries for Musicology, DLfM@JCDL 2014, London, United Kingdom, September 12, 2014, pp. 1–8. ACM. https://doi.org/10.1145/2660168.2660172 (2014)
11. Burstyn, S.: In quest of the period ear. Early Music **25**(4), 692–701 (1997)
12. Cheng, J., Teevan, J., Iqbal, S.T., Bernstein, M.S.: Break it down: A comparison of macro- and microtasks. In: B. Begole, J. Kim, K. Inkpen, W. Woo (eds.) Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18–23, 2015, pp. 4061–4064. ACM. https://doi.org/10.1145/2702123.2702146 (2015)
13. Cox, S., Little, C.: Time ontology in OWL. W3C Working Draft 12 July 2016, World Wide Web Consortium (W3C). http://www.w3.org/TR/owl-time/ (2016)
14. Cunningham, S.J., Nichols, D.M., Bainbridge, D., Ali, H.: Social music in cars. In: H. Wang, Y. Yang, J.H. Lee (eds.) Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27–31, 2014, pp. 457–462. http://www.terasoft.com.tw/conf/ismir2014/proceedings/T083_184_Paper.pdf (2014)
15. Fields, B., Rhodes, C.: Listen to me—don't listen to me: what communities of critics tell us about music. In: Mandel et al. [26], pp. 199–205. https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/173_Paper.pdf
16. Golden, P., Shaw, R.B.: Nanopublication beyond the sciences: the PeriodO period gazetteer. PeerJ Comput. Sci. **2**, e44 (2016). https://doi.org/10.7717/peerj-cs.44
17. Gouyon, F., Herrera, P., Martins, L.G., Müller, M. (eds.): Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S. Bento Da Vitória, Porto, Portugal, October 8–12, 2012. FEUP Edições (2012)
18. Halpin, H., Presutti, V.: The identity of resources on the web: an ontology for web architecture. Appl. Ontol. **6**(3), 263–293 (2011). https://doi.org/10.3233/AO-2011-0095
19. Herbert, T., Barlow, H.: Music & The British Military in the Long Nineteenth Century. Oxford University Press, Oxford. https://books.google.co.uk/books?id=nr3ANASf_ywC (2013)
20. Heywood, I.: Sensory Arts and Design. Sensory Studies Series. Bloomsbury Publishing, London. https://books.google.co.uk/books?id=XcF-DQAAQBAJ (2017)
21. Juslin, P.N., Västfjäll, D.: Emotional responses to music: the need to consider underlying mechanisms. Behav. Brain Sci. **31**, 559–575 (2008). https://doi.org/10.1017/S0140525X08005293
22. Kamalzadeh, M., Baur, D., Möller, T.: A survey on music listening and management behaviours. In: Gouyon et al. [17], pp. 373–378. http://ismir2012.ismir.net/event/papers/373-ismir-2012.pdf
23. Larkou, G., Metochi, J., Chatzimilioudis, G., Zeinalipour-Yazti, D.: CLODA: A crowdsourced linked open data architecture. In: MDM (2), pp. 104–109. IEEE (2013)
24. Library of Congress: Extended date/time format (EDTF) 1.0. Draft submission 13 January 2012, Library of Congress. http://www.loc.gov/standards/datetime/pre-submission.html (2012)
25. Dearn, L., Gross, J., Price, S., Pitts, S.: The listening experience of the classical concert hall: the value of qualitative research with current audiences. In: Barlow and Rowland [3]. http://ledbooks.org/proceedings2017/#sec_115_h1
26. Mandel, M.I., Devaney, J., Turnbull, D., Tzanetakis, G. (eds.): Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7–11, 2016 (2016)
27. Page, C.: Listening practice—an introduction. Early Music **XXV**(4), 591–592 (1997)
28. Pearson, I.E.: Listening and performing: experiences of twentieth-century British wind players. In: Barlow and Rowland [3]. http://ledbooks.org/proceedings2017#sec_227_h1
29. Raimond, Y., Abdallah, S.A., Sandler, M.B., Giasson, F.: The music ontology. In: S. Dixon, D. Bainbridge, R. Typke (eds.) Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23–27, 2007, pp. 417–422. Austrian Computer Society. http://ismir2007.ismir.net/proceedings/ISMIR2007_p417_raimond.pdf (2007)
30. Roengsamut, B., Kuwabara, K.: Interactive refinement of linked data: toward a crowdsourcing approach. In: N.T. Nguyen, B. Trawinski, R. Kosala (eds.) Intelligent Information and Database Systems—7th Asian Conference, ACIIDS 2015, Bali, Indonesia, March 23–25, 2015, Proceedings, Part I, *Lecture Notes in Computer Science*, vol. 9011, pp. 3–12. Springer. https://doi.org/10.1007/978-3-319-15702-3_1 (2015)
31. Schäfer, T., Sedlmeier, P., Städtler, C., Huron, D.: The psychological functions of music listening. Front. Psychol. (2013). https://doi.org/10.3389/fpsyg.2013.00511
32. Schedl, M., Eghbal-Zadeh, H., Gómez, E., Tkalcic, M.: An analysis of agreement in classical music perception and its relationship to listener characteristics. In: Mandel et al. [26], pp. 578–583. https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/260_Paper.pdf
33. Schellenberg, E.G.: Cognitive performance after listening to music: A review of the Mozart effect. In: R. MacDonald, G. Kreutz, L. Mitchell (eds.) Music, Health, and Wellbeing. Oxford University Press, Oxford. https://doi.org/10.1093/acprof:oso/9780199586974.003.0022 (2012)
34. Simperl, E., Acosta, M., Norton, B.: A semantically enabled architecture for crowdsourced linked data management. In: Baeza-Yates, R.A., Ceri, S., Fraternali, P., Giunchiglia, F. (eds.) CrowdSearch, *CEUR Workshop Proceedings*, vol. 842, pp. 9–14. CEUR-WS.org. (2012)
35. Summers, C., Popp, P.: Temporal music context identification with user listening data. In: M. Müller, F. Wiering (eds.) Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26–30, 2015, pp. 59–64. http://ismir2015.uma.es/articles/195_Paper.pdf (2015)
36. W3C OWL Working Group: OWL 2 Web Ontology Language: Document overview (second edition). W3C Recommendation 11 December 2012, World Wide Web Consortium (W3C). https://www.w3.org/TR/owl2-overview/ (2012)
37. Watson, D., Mandryk, R.L.: Modeling musical mood from audio features and listening context on an in-situ data set. In: Gouyon et al. [17], pp. 31–36. http://ismir2012.ismir.net/event/papers/031-ismir-2012.pdf
38. Wegman, R.C.: Music as heard: listeners and listening in Late-Medieval and Early Modern Europe (1300–1600). Musical Q. **82**(3–4), 432–433 (1998)
39. Wiesecke, J.: Samuel Pepys and his experiences of music at Restoration theatres. In: Barlow and Rowland [3]. http://ledbooks.org/proceedings2017/#sec_245_h1